مجلة دجلة • المجلد ٦ ،العدد ٤ (٢٠٢٣)، ص: ٦٨ - ٨٠
ISSN: 2222-6583

Securing the Future: Concise Strategies for Mitigating Security and Control Threats Accompanying the AI Revolution

Hussein Abdulkhaleq Saleh
Directorate General of Education in Dhi Qar, Nasiriyah, Dhi Qar, Iraq.
Author E-mail: hussein.abd.alkhaliq@gmail.com
ORCID: 0009-0003-3426-7168

Abstract:

Artificial intelligence (AI) is rapidly advancing and being integrated into many aspects of society. While AI enables numerous benefits, it also introduces new security and control risks. This study explores strategies for mitigating security threats posed by the ongoing AI revolution. The key areas of concern include data poisoning, adversarial attacks, AI algorithm biases, automation-related unemployment, and artificial general intelligence. The potential solutions evaluated include improved training data curation, adversarial training, algorithmic fairness techniques, policy interventions for job displacement, and AI safety research. A multi-pronged approach is recommended that focuses on technological innovations, workforce development programs, regulatory oversight, and research to ensure aligned and controllable advanced AI systems. Proactive collaboration among stakeholders in industry, government, and academia is crucial for positive security outcomes as AI becomes further entrenched. The recommendations outlined in this paper promote beneficial uses of AI while better understanding and reducing attendant risks.

Keywords: artificial intelligence security, AI control safety, AI revolution threats

1- Introduction:

Artificial intelligence (AI) can be defined as computational systems that perform tasks normally requiring human cognition and expertise [1]. It refers to the simulation of human intelligence by machines that are programmed to think like humans and mimic their actions. It involves comprehensive thinking and many other actions, options, and outcomes beyond simple imitation [2]

AI is a multidisciplinary field, focused on automating tasks that require human intelligence. AI is revolutionizing all aspects of life and can be used as a tool in many fields to rethink how we combine data, analyze it, and make decisions [3].

Artificial intelligence collects, processes, and learns from data to perform automated tasks and decision-making optimally and efficiently. It is ubiquitous in today's society, silently collecting data and optimizing tasks and decision-making as computers are trained to mimic how humans act, sense, and think [4].

One of the important branches of artificial intelligence is machine learning, which plays a crucial role in advancing the capabilities of artificial intelligence and has numerous applications in various fields. It focuses on training machines to learn from existing data and make predictions and allows computers to imitate human behaviors and accomplish specific tasks [5].

With advances in machine learning, AI is achieving new milestones in capabilities such as computer vision, speech recognition, and strategic gameplay. In addition, AI systems are being embedded in vital areas, such as social media services, autonomous vehicles, predictive policing, personalized medicine, and other domains. While this new technology wave promises to transform society in positive ways, some experts warn that may also introduce novel security threats [6] [7] [8].

As increasing amounts of economic activity and infrastructure depend on AI, new attack vectors and points of failure are introduced, where malicious actors may attempt to subvert, mislead, or otherwise exploit AI systems. For instance, biases encoded in algorithms can lead to unintended discrimination and injustice or impair the AI model algorithm's function by attacking the training datasets.

On the other hand, the rapid automation of jobs can significantly disrupt labor markets and destabilize economies and societies [9][10][11]. AI and automation technologies have already disrupted various sectors of the economy, particularly routine and repetitive tasks [12]. This raises concerns about mass technological unemployment and growing economic inequality, where developing countries face challenges in managing the impacts of automation, requiring strategies such as upskilling, safety nets, and investments in labor-heavy sectors.

Building ever more intelligent and powerful AI systems could eventually result in systems that are more powerful than humans, creating existential risks [13], where the risks of hypothetical future AI systems with general cognitive abilities surpassing humans remain highly speculative, but potentially extreme risks. Thus, there is a pressing need for a systematic discussion and illustration of these risks to better inform efforts to mitigate them [14].

AI is revolutionizing sectors from healthcare to transportation by automating tasks and augmenting human capabilities. As applications proliferate, AI promises unprecedented benefits such as personalized medicine, autonomous logistics, and advanced education. However, responsible development requires proactively addressing attendant security, economic, and social risks to maximize positive outcomes.

This paper discusses pressing security and control issues surrounding current and foreseeable real-world applications of AI, as well as long-term concerns and suggests strategies for mitigating risks while allowing continued innovation and adoption of this transformative technology.

2- Literature Review

Artificial intelligence (AI) has revolutionized numerous sectors, including healthcare, finance, transportation, and defense. However, the rapid advancement of AI technologies has also introduced new security and control threats.

- Security Threats in AI:

Several studies have underscored the potential security risks associated with AI. Brundage et al. [15] discussed the potential misuse of AI and its implications for national security, suggesting that AI could be used to automate tasks involved in cyber-attacks, thereby increasing their scale and speed. similarly, Sharif et al. [16], have shown how machine learning models can be manipulated through adversarial attacks, leading to incorrect outputs. Hu [17] discusses different security threats and provides countermeasures for each stage of the AI lifecycle.

- Control Threats in AI:

The control problem in AI is another area of concern. Russell et al. [18] argue that as AI systems become more autonomous, there is a growing risk of these systems deviating from their intended functions. This could lead to undesirable outcomes if not properly managed. Hong [19] proposed an artificial intelligence-based security control platform that uses self-learning and monitoring to detect and block attacks.

- Mitigation Strategies:

In response to these threats, researchers have proposed various mitigation strategies. Liu et al. [20] proposed a framework for defending against adversarial attacks in AI, while Hadfield-Menell et al. [21] suggested a value alignment approach to address the control problem in AI. Bertino [22] focuses on the adoption of AI techniques in the security industry and the rise of adversarial AI, where AI systems are subverted for malicious purposes.

In conclusion, while AI presents numerous opportunities, it is crucial to address the accompanying security and control threats to ensure its safe and beneficial use. This study contributes to this ongoing discourse by proposing novel mitigation strategies.

3- Concerns and Threats:

Artificial intelligence as a general-purpose tool has become a great force globally in the past decade, it has enabled many important applications across all sectors of society, such as recognizing pathological medical images and translating text from many languages. However, the enormous potential for innovation and technological advances and the chances that AI systems provide come with hazards and risks, some of which are not yet fully explored, let alone fully understood. In the next subsections, the important concerns and threats will be listed, followed by the proposed solutions.

Data poisoning attacks

Many modern AI systems rely on training datasets to learn to perform tasks. These datasets may contain errors, omissions, or purposefully introduced misinformation that is incorporated into AI models' algorithms, impairing their function. For example, an autonomous vehicle's object recognition system could be compromised if its training images are poisoned with mislabeled data [23]. Similarly, poisoning the training data for AI-assisted cyber intrusion detection systems may blind them to certain attacks [24]. Strategies to increase training data integrity include the following:

- Curating datasets with redundancy: Having multiple redundant copies of the training data can help identify when some copies have been altered or poisoned.
- Provenance-tracking: Recording the history and origin of the training data (its provenance) can help spot data that comes from suspicious or tampered sources.
- Testing for statistical anomalies: Analyzing the statistics of the training data and looking for anomalies can reveal if some subset of the data has different characteristics due to poisoning.

Adversarial Examples

Adversaries can craft input data in the form of adversarial examples, which are inputs intentionally designed to appear normal and legitimate to human observers but fool AI systems into making incorrect decisions by causing machine learning models to make mistakes.

For example, images altered in subtle ways can dupe computer vision algorithms into misclassifying objects [25], where an adversary could take an image of a dog and alter the pixel values in subtle ways imperceptible to humans, crafting an "adversarial image" that gets misclassified as a cat by a machine learning model despite still appearing as a dog to people.

Similarly, speech recognition systems struggle to detect garbled audio adversarial attacks [26], where adversaries can add small perturbations to audio signals that humans hear as normal speech, but AIs transcribe incorrectly.

Some defenses against adversarial examples include the following:

- Retraining models on adversarial examples: Exposes the model to more challenging cases which makes it more robust.
- Architectures with less sensitivity to small perturbations: Designing models such that small input tweaks do not drastically alter outputs.
- Detecting adversarial inputs: Building systems to recognize when inputs have unnatural statistical properties.

Algorithmic Biases

AI systems trained on flawed, limited, or skewed data can exhibit biases that are amplified in real-world applications, where algorithms encoding racial, gender, or ideological prejudices could further marginalize groups when used in areas such as insurance, lending, policing, and social media [27].

In the financial sector, algorithmic biases can influence decisions related to loan approvals, credit limits, and credit score estimations, leading to unjust outcomes for consumers [28]. In the realm of autonomous vehicles, biases present in algorithms predicting pedestrian trajectories can elevate the risks for vulnerable pedestrians, including individuals with disabilities, the elderly, and children. Furthermore, biases within machine learning algorithms employed in industrial and safety-critical applications, especially those reliant on complex inputs such as images, can yield substantial consequences [29]

To address algorithmic biases, some techniques try to reduce unwanted biases, as follows:

- Testing algorithms across different demographic groups to detect disparities in performance.
- Modifying the training process to ensure that algorithms perform equally well for different groups.
- Using techniques such as debiasing to remove unwanted correlations from datasets.
- Design algorithms that are inherently fairer, interpretable, and transparent.

Automation Unemployment

In recent years, AI has been automating an expanding range of occupations, thereby displacing some human work activities. For example, manufacturing robots have replaced some factory jobs such as assembly lines.

If we assume that AI automation does lead to significant technological unemployment, it could increase economic inequality as the benefits flow to a smaller group able to afford the new technologies.

However, predictions vary on whether AI automation will lead to large-scale technological unemployment, where large numbers of people struggle to find new jobs after being displaced.

Some studies predict that huge numbers of jobs will be automated by AI [30]. Others argue that we will continue to adapt and find new types of work, just as happened in past economic revolutions [31].

The impact of AI on jobs remains an open question and an active area of research; however, mitigating the potential impacts is a required proactive action, as included in the following proposals:

- Education reform to teach new skills needed for emerging jobs.
- Retraining programs for displaced workers.
- Basic income guarantees.
- Tax incentives or subsidies to encourage job creation in new fields.

Artificial General Intelligence

AGI refers to hypothetical future AI systems that have general cognitive abilities at or above the human level. Unlike current narrow AI systems that are specialized for specific tasks, AGI would be able to reason, plan, communicate, perceive, and learn across a broad range of domains [32].

AGI aims to replicate human intelligence through computer systems and has gained recognition as a future technology [33], where it can perform tasks that require human-level intelligence, such as reasoning, problem-solving, decision-making, and understanding human emotions and social interactions [34].

While general-human-level AGI may not be imminent, some researchers recommend early safety research and ethics training for advanced AI [35] [36].

The development of AGI systems presents complex security and control challenges that arise from the uncertainty surrounding the properties, capabilities, and motivations of future AGIs, including the following:

1- Uncertainty over impacts: Since hyperintelligent systems are unprecedented, their full capabilities and consequences are difficult to predict and control [37]. For instance, in the education field, AGI can perform tasks that require human-level intelligence, such as reasoning, problem-solving, decision-making, and understanding human emotions and social interactions [38] [39]. However, there is uncertainty over the impacts of AGI. It is important to consider the ethical issues in education faced by AGI and how it will affect human educators [40]

2- Indifference to human values: Advanced AGI may not inherently share human morals, preferences, and values. Without explicit programming of human values, its objectives could be misaligned with ours [41]. The sensitivity and amount of data being collected through AI advancements have prompted scholars in the neuroethics community to identify this issue [42]. Aligning the goals of hyperintelligent machines with human values is crucial for safety in AGI systems, but human values are complex and cannot be easily formalized [43]. This concern is important because AI systems can produce harmful inequalities with various manifestations [44].

3- Comprehensibility: Highly intelligent systems may become too complex for humans to fully understand, making oversight difficult [45]. The lack of explainability and comprehensibility in advanced AI systems can hinder users' ability to comprehend the decisions made by these systems. Subsequently, this leads to a lack of trust in safety-critical applications [46][47].

To address these risks, the recommended initiatives include the following:

- Value alignment: Ensure AI goals and values align with human values [48] [49][50]
- AI safety: Design theoretically robust goals and safeguards [51].
- Containment strategies: Theoretical ways to control superintelligent systems [52] [53][54].

4- General mitigation strategies:

AI security risks can be mitigated through various strategies. One approach is to examine privacy risks throughout the AI life cycle and implement privacy-enhancing solutions [55].

Another strategy is to consider the dimensions of AI risk, including military, political, economic, social, environmental, psychophysiological, and spiritual, and propose alternatives within a democratic governance framework [56].

Additionally, securing the big data infrastructure using intelligent agent paradigms and big data techniques can enhance data security [57].

Furthermore, confining AI systems to specific tasks, such as Oracle AIs that can only answer questions, can provide a safer approach to AI development [58].

In addition, the following combination of strategies is recommended across the domains of technology, policy, and research to tackle interrelated AI threats while sustaining innovation.

- Improve technical protections, testing, and monitoring to make AI systems more secure, transparent, and accountable from the outset.
- use best practices in cybersecurity and software engineering for AI design.
- Support displaced workers through educational opportunities and social programs to ease workforce transitions and provide alternative economic roles.
- Enact regulations and standards to safeguard the public, while allowing the flexibility for ongoing advances in AI capabilities. Policymakers, ethicists, engineers, and users should collaborate to shape appropriate regulations and laws.
- Expand public and private funding for focused research initiatives into promising approaches for making emerging AI technologies more secure, trustworthy, and aligned with human values over the long term.
- Cooperative proactive efforts between stakeholders in companies developing AI, government agencies overseeing impacts, and researchers exploring solutions. With vigilance and foresight, society can maximize benefits and minimize risks as this technological force continues to advance.

5- Using AI to secure AI

AI itself can be a powerful tool to mitigate these threats and secure AI applications. This potential, along with the associated challenges and future directions, will be explored later in this section.

- AI for Security: AI can be leveraged to improve the security of other AI systems. Machine learning algorithms can be used to detect anomalies or suspicious activities in AI applications, thereby enhancing their security [59]. These algorithms can learn from historical data and make predictions about potential security threats, enabling proactive security measures [60]. This approach allows for the early detection and mitigation of security threats, reducing the potential damage caused by these threats [59].
- AI for Control: AI can also be used to ensure the proper control of other AI systems. Techniques such as reinforcement learning can be used to train AI

systems to behave in a certain manner, thereby reducing the risk of these systems deviating from their intended functions. This can help ensure that AI systems perform their tasks as expected, thereby reducing the risk of unexpected and potentially harmful behavior [61].

- AI and Blockchain: The integration of AI with other technologies such as blockchain can further enhance the security of AI applications. Blockchain can ensure the integrity of shared data or models used by AI, including deep learning and many machine learning techniques. This can help prevent data tampering and ensure the reliability of AI applications [62].
- Challenges and Future Directions: While the use of AI to secure AI applications holds promise, it also presents challenges. These include the risk of adversarial attacks on the AI systems used for security and the need for transparency and interpretability in these systems. Future research should focus on addressing these challenges and exploring new ways to use AI to enhance the security and control of AI applications.

6- Conclusions:

This paper provides an overview of key near and long-term security threats associated with artificial intelligence systems, including data poisoning, adversarial examples, algorithmic biases, employment impacts of automation, AGI, and general mitigation strategies. A range of technological, policy, economic, and other interventions have been proposed to address these multifaceted challenges.

The potential benefits of AI are enormous, but the large-scale adoption of AI technologies also brings new and unforeseen threats. Further research is needed to overcome these challenges and ensure the effectiveness of AI applications from a security perspective.

In addition, this paper has touched upon the concept of using AI to enhance the security of AI applications. AI can serve as a potent instrument to counteract these threats and safeguard AI applications. This potential, along with the associated challenges and future directions, has been explored.

Protecting civilization from the adverse consequences of AI transformative technologies while allowing ongoing progress requires prudent management of risks. AI security strategies should emphasize defensive measures and research into the value alignment and controllability of future AI. By ensuring the safety, traceability, transparency, and explainability of AI applications, stakeholders can develop trustworthy and ethically reliable AI that aligns with human values.

Through coordinated efforts among industry, government, and academia, solutions may be achieved to balance the benefits of AI with the need for security. This approach, coupled with the potential of using AI to secure AI applications, will help governments and policymakers secure humanity’s future in the age of intelligent machines.

References

[1] Tugce Gokdeniz, S., Buyuksungur, A., & Eray Kolsuz, M. (2023). Artificial Intelligence in Dentistry. IntechOpen. doi: 10.5772/intechopen.111532.

[2] Yarali, A. (2023). Artificial Intelligence. In From 5G to 6G, A. Yarali (Ed.). https://doi.org/10.1002/9781119883111.ch9

[3] Noble, R., & Noble, D. (2023). Artificial Intelligence. In Understanding Living Systems, Understanding Life (pp. 99–112). chapter, Cambridge: Cambridge University Press.

[4] Duan, F.L. (2023). Artificial Intelligence. In: When AIAA Meets IEEE. Springer, Singapore. https://doi.org/10.1007/978-981-19-8394-8_2

[5] Abdulmajeed, I., Nassreddine, G., El Arid, A. A., & Younis, J. (2023). Machine Learning Approach in Human Resources Department. In S. Kaddoura (Ed.), Handbook of Research on AI Methods and Applications in Computer Engineering (pp. 271-294). IGI Global. https://doi.org/10.4018/978-1-6684-6937-8.ch013

[6] Qi, W., Pan, J., Lyu, H., & Luo, J. (2023). Excitements and Concerns in the Post-ChatGPT Era: Deciphering Public Perception of AI through Social Media Analysis. arXiv preprint arXiv:2307.05809.

[7] Tidjon, L. N., & Khomh, F. (2022). Never trust, always verify: a roadmap for Trustworthy AI?. arXiv preprint arXiv:2206.11981.

[8] Mishra, S. (2022). Artificial intelligence: A review of progress and prospects in medicine and healthcare. Journal of Electronics, Electromedical Engineering, and Medical Informatics, 4(1), 1-23.

[9] Mohd., Faishal., Saju, John, Mathew., Khriemenuo, Pusa. (2023). The future of work: AI, automation, and the changing dynamics of developed economies. World Journal Of Advanced Research and Reviews, 18(3):620-629. doi: 10.30574/wjarr.2023.18.3.1086

[10] Jeremy, Schulz. (2022). Future Shocks: Automation Meets the Pandemic. American Behavioral Scientist, 000276422211272-000276422211272. doi: 10.1177/00027642221127235

[11] Lukas, Schlogl., Andy, Sumner. (2020). Automation, Politics, and Public Policy. 79-83. doi: 10.1007/978-3-030-30131-6_6

[12] Rolf, Clauberg. (2020). Challenges of digitalization and artificial intelligence for modern economies, societies and management. RUDN Journal of Economics, 28(3):556-567. doi: 10.22363/2313-2329-2020-28-3-556-567

[13] Hendrycks, D., & Mazeika, M. (2022). X-risk analysis for ai research. arXiv preprint arXiv:2206.05862.

[14] Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks. arXiv preprint arXiv:2306.12001.

[15] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

[16] Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016, October). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In

Proceedings of the 2016 acm sigsac conference on computer and communications security (pp. 1528-1540).

[17] Hu, Y., Kuang, W., Qin, Z., Li, K., Zhang, J., Gao, Y., ... & Li, K. (2021). Artificial intelligence security: Threats and countermeasures. ACM Computing Surveys (CSUR), 55(1), 1-36.

[18] Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.

[19] Hong, J. H., & Lee, B. Y. (2021). Artificial Intelligence-based Security Control Construction and Countermeasures. The Journal of the Korea Contents Association, 21(1), 531-540.

[20] Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. (2018). Defense against adversarial attacks using high-level representation guided denoiser. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1778-1787).

[21] Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. Advances in neural information processing systems, 29.

[22] Bertino, E., Kantarcioglu, M., Akcora, C. G., Samtani, S., Mittal, S., & Gupta, M. (2021, April). AI for Security and Security for AI. In Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy (pp. 333-334).

[23] Sabharwal, K., Kabir, M., & Chatterjee, K. (2022). Adversarial Image Perturbation: A Survey.

[24] Kravchik, M., Biggio, B., & Shabtai, A. (2021, March). Poisoning attacks on cyber attack detectors for industrial control systems. In Proceedings of the 36th Annual ACM Symposium on Applied Computing (pp. 116-125).

[25] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[26] Carlini, N., & Wagner, D. (2018, May). Audio adversarial examples: Targeted attacks on speech-to-text. In 2018 IEEE security and privacy workshops (SPW) (pp. 1-7). IEEE.

[27] Cave, S., & ÓhÉigeartaigh, S. S. (2018, December). An AI race for strategic advantage: rhetoric and risks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 36-40).

[28] Bae, A., & Xu, S. (2022, November). Discovering and Understanding Algorithmic Biases in Autonomous Pedestrian Trajectory Predictions. In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (pp. 1155-1161).

[29] Risser, L., Picard, A., Hervier, L., & Loubes, J. M. (2022). A survey of Identification and mitigation of Machine Learning algorithmic biases in Image Analysis. arXiv preprint arXiv:2210.04491.

[30] Winick, E. (2018, January 25). Every study we could find on what automation will do to jobs, in one chart. MIT Technology Review. Retrieved March 3, 2023, from https://www.technologyreview.com/2018/01/25/146020/every-study-we-could-find-on-what-automation-will-do-to-jobs-in-one-chart/

[31] AGENDA, D. (2020, October 26). Don’t fear AI. The tech will lead to long-term job growth. World Economic Forum. Retrieved April 3, 2023, from https://www.weforum.org/agenda/2020/10/dont-fear-ai-it-will-lead-to-long-term-job-growth/

[32] Zhao, L., Zhang, L., Wu, Z., Chen, Y., Dai, H., Yu, X., ... & Liu, T. (2023). When brain-inspired ai meets agi. Meta-Radiology, 100005.

[33] Guerrero, J. M. (2023). Artificial general intelligence. In Mind Mapping and Artificial IntelligenceeBooks (pp. 181–195). https://doi.org/10.1016/b978-0-12-820119-0.00009-1

[34] Aditi, Guha, Choudhury. (2023). Artificial Energy General Intelligence AEGI. doi: 10.31219/osf.io/ye254

[35] Catalin, Mitelut., Ben, Smith., Peter, Vamplew. (2023). Intent-aligned AI systems deplete human agency: the need for agency foundations research in AI safety. arXiv.org, abs/2305.19223 doi: 10.48550/arXiv.2305.19223

[36] Jonas, Schuett., Noemi, Dreksler., Markus, Anderljung., David, McCaffary., Lennart, Heim., Emma, Bluemke., Ben, Garfinkel. (2023). Towards best practices in AGI safety and governance: A survey of expert opinion. arXiv.org, abs/2305.07153 doi: 10.48550/arXiv.2305.07153

[37] Craig, Van, Slyke., Richard, L., Johnson., Jalal, Sarabadani. (2023). Generative Artificial Intelligence in Information Systems Education: Challenges, Consequences, and Responses. Communications of The Ais, 53(1):1-21. doi: 10.17705/1cais.05301

[38] Rothman, D. (2018). Artificial intelligence by example: develop machine intelligence from scratch using real artificial intelligence use cases. Packt Publishing Ltd.

[39] Sathian, Dananjayan., Gerard, Marshall, Raj. (2020). Artificial Intelligence during a pandemic: The COVID-19 example.. International Journal of Health Planning and Management, 35(5):1260-1262. doi: 10.1002/HPM.2987

[40] Ehsan, Latif., Gengchen, Mai., Matthew, Nyaaba., Xuansheng, Wu., Ninghao, Liu., Guoyu, Lu., Sheng, Li., Tianming, Liu., Xiaoming, Zhai. (2023). Artificial General Intelligence (AGI) for Education. arXiv.org, abs/2304.12479 doi: 10.48550/arXiv.2304.12479

[41] Conn, A. (2015, November 14). Benefits and Risks of artificial intelligence. Future of Life Institute. Retrieved March 3, 2023, from https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1

[42] Daniel, Lim. (2019). Divergent Values and Adaptive Preferences: A Chinese Challenge?. Ajob Neuroscience, 10(3):132-134. doi: 10.1080/21507740.2019.1632962

[43] Devillers, L., Fogelman-Soulié, F., & Baeza-Yates, R. (2021). AI & Human Values: Inequalities, Biases, Fairness, Nudge, and Feedback Loops. Reflections on Artificial Intelligence for Humanity, 76-89.

[44] Lawrence, M., Lesser. (1998). Countering Indifference Using Counterintuitive Examples. Teaching Statistics, 20(1):10-12. doi: 10.1111/J.1467-9639.1998.TB00750.X

[45] Yampolskiy, R. V. (2019). Unexplainability and incomprehensibility of artificial intelligence. arXiv preprint arXiv:1907.03869.

[46] Sado, F., Loo, C. K., Liew, W. S., Kerzel, M., & Wermter, S. (2023). Explainable Goal-driven Agents and Robots-A Comprehensive Review. ACM Computing Surveys, 55(10), 1-41.

[47] Thao, Minh, Phuong, Ngo., Nicole, C., Krämer. (2022). I humanize, therefore I understand? Effects of explanations and humanization of intelligent systems on perceived and objective user understanding. doi: 10.31234/osf.io/6az2h

[48] Nay, J., & Daily, J. (2022). Aligning Artificial Intelligence with Humans through Public Policy. arXiv preprint arXiv:2207.01497.

[49] Mechergui, M., & Sreedharan, S. (2023). Goal Alignment: A Human-Aware Account of Value Alignment Problem. arXiv preprint arXiv:2302.00813.

[50] Betty, Hou., Brian, Patrick, Green. (2023). A Multi-Level Framework for the AI Alignment Problem. arXiv.org, abs/2301.03740 doi: 10.48550/arXiv.2301.03740

[51] Soares, N., & Fallenstein, B. (2017). Agent foundations for aligning machine intelligence with human interests: a technical research agenda. The technological singularity: Managing the journey, 103-125.

[52] Avinash, Kr., Dubey., Dwaipayan, Mukherjee. (2022). Containment control of heterogeneous multi-agent systems. Advances in Control and Optimization of Dynamical Systems, 55(22):363-368. doi: 10.1016/j.ifacol.2023.03.061

[53] Qin, Fu., Pengfei, Yu., Guangzhao, Xu., Jianrong, Wu. (2019). Containment control for partial differential multi-agent systems. Physica A-statistical Mechanics and Its Applications, 529:121549-. doi: 10.1016/J.PHYSA.2019.121549

[54] Chen, Yuan., Huaicheng, Yan., Yuan, Wang., Yufang, Chang., Xisheng, Zhan. (2022). Formation-containment control of heterogeneous linear multi-agent systems with adaptive event-triggered strategies. International Journal of Systems Science, 53(9):1942-1957. doi: 10.1080/00207721.2022.2031339

[55] Shahriar, S., Allana, S., Fard, M. H., & Dara, R. (2023). A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle. IEEE Access.

[56] Colin, Garvey. (2018). AI Risk Mitigation Through Democratic Governance: Introducing the 7-Dimensional AI Risk Horizon. 366-367. doi: 10.1145/3278721.3278801

[57] Mihai, Horia, Zaharia. (2019). Using Intelligent Agents Paradigm in Big Data Security Risks Mitigation. 76-97. doi: 10.4018/978-1-5225-7277-0.CH005

[58] Armstrong, S. (2013). Risks and Mitigation Strategies for Oracle AI (pp. 335-347). Springer Berlin Heidelberg.

[59] Buehrer, G. (2023, November 6). Building for the future: The enterprise generative AI application lifecycle with Azure AI. Azure. Retrieved November 13, 2023, from https://azure.microsoft.com/en-us/blog/building-for-the-future-the-enterprise-generative-ai-application-lifecycle-with-azure-ai/

[60] Maraju, K. (2018, January 1). Applying AI in application Security. ISACA. Retrieved April 5, 2023, from https://www.isaca.org/resources/isaca-journal/issues/2018/volume-1/applying-ai-in-application-security

[61] Kumar, R. S. S. (2021, December 9). Best practices for AI security risk management. Microsoft Security Blog. Retrieved February 4, 2023, from https://www.microsoft.com/en-us/security/blog/2021/12/09/best-practices-for-ai-security-risk-management/

[62] R. Tian, L. Kong, X. Min and Y. Qu, "Blockchain for AI: A Disruptive Integration," 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hangzhou, China, 2022, pp. 938-943, doi: 10.1109/CSCWD54268.2022.9776023.